LAMOST 异构环境信息检索系统的设计与实现*

王政[1,3]，王锋[2,3]，田园[1]，李建[1]，赵永恒[1,3]

（1. 中国科学院光学天文重点实验室，国家天文台，北京 100101；

2. 广州大学天体物理中心，广东 广州 510006；3. 中国科学院大学，北京 100049）

摘要：环境信息是大天区面积多目标光纤光谱望远镜（LAMOST，即郭守敬望远镜）运行状态的一部分，用于辅助望远镜的日常维护、观测以及后期数据处理。环境信息来源于多个子系统，而子系统的数据存储由不同单位设计，使得数据存储环境复杂，不利于统一检索。基于异步协程，我们提出了一种异构环境信息检索的方法。该方法利用数据库代理对象工厂整合多个数据库及检索字段信息，通过客户端和 Web 浏览器提供远程服务，以自定义检索命令的方式实现对异构数据库的检索。本方法充分考虑了后期升级的需求，保留多种数据库和应用接口，便于更多数据库和应用的接入。目前基于该方法开发的检索系统已部署于 LAMOST 运行环境，在实际运行中简化了运行环境检索操作，提高了工作效率和数据获取的精准度，提升了维护和观测效率。

关键词：信息检索；异步协程；LAMOST

中图分类号: TP311　文献标识码: A

LAMOST 是我国自主研制的大口径兼具大视场的反射式施密特光学望远镜，安置于中国科学院国家天文台兴隆观测站。该望远镜可以同时观测 4000 个目标，是世界上单次观测获取光谱数最多的望远镜[1]。到目前为止，LAMOST 已观测、生产、发布了超过 1000 万条天文光谱①。

除了生产天文光谱之外，LAMOST 各子系统每天都会生成数百万条实时状态信息，并存储在不同类型的数据库中[2]。如图 1 所示，LAMOST 运行环境是多个子系统状态信息数据的集合，提供各种环境信息，包含 Weather 子系统中的气象信息和 DIMM 信息、TCS 子系统中的主动光学和机架风速信息、GUID 系统中的导星信息、ICS 子系统中的相机温度信息、OSS 系统中的观测目标信息以及 OCS 系统的观测信息，是一种由 MySQL、SQLite、PostgreSQL 多种数据库组成的异构存储系统[3-4]。

* 基金项目: 国家自然科学基金（11703044）资助.
收稿日期: 2020－04－21；修订日期:
作者简介: 王政，男，博士. 研究方向：大型望远镜观测控制技术、FPGA 设计. Email: wzheng@bao.ac.cn

① http://dr7.lamost.org/

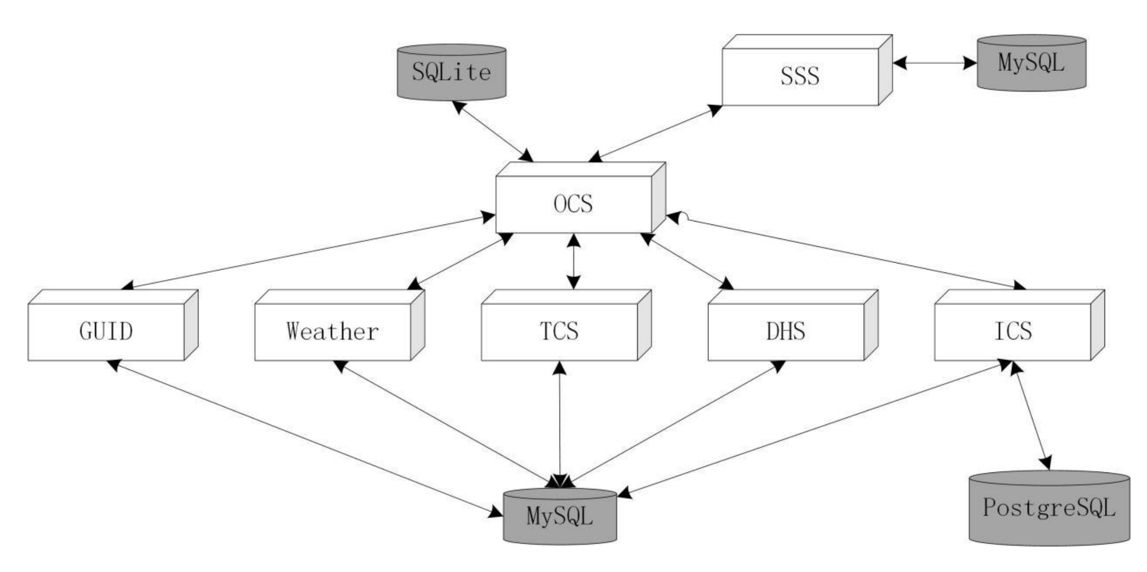


图 1 LAMOST 环境信息存储结构图

Figure 1 The storage structure of environmental information

环境信息为望远镜的观测、维护以及后期数据处理提供重要的参考信息。根据信息的应用和重要性，可分以下几种类型：

a. 致命提醒：会影响望远镜安全的突发事件。观测时，如遇湿度过大、大风等突发状况，以声音、醒目频闪颜色等方式提醒观测者立刻停止观测，关闭望远镜，以保护望远镜安全。

b. 辅助观测：视觉显示观测天区和观测目标的关系，辅助驻站天文学家选择最佳的观测目标。

c. 故障报警：望远镜子系统故障提醒信息。观测时，提醒观测者故障位置，便于即使维修。在日常维护中，现场维护人员需快速检索望远镜各子系统的状态信息，如有异常，立刻锁定问题，及时维修，确保望远镜处于最佳运行状态。非现场维护人员通过远程检索各子系统的状态信息，协助现场工作人员维护望远镜。维护人员通过统计长期环境信息，结合维护日志以及观测中遇到的问题，提出望远镜维护方案，降低维护成本和时间，提高观测效率。

d. 问题回溯：在后期数据处理过程中，数据处理人员对争议光谱可回溯观测时环境信息，确定问题导向。

e. 环境展示：通过大屏幕集中展示当前观测及运行状态。

当前，LAMOST 运行环境中的各种状态信息相互孤立，只能在局域网内以复杂的检索语句对同一种数据库进行检索，非现场工作人员只能通过多个代理跳转入内网才可以检索信息，操作复杂，且存在暴露内网的安全隐患。检索语句因数据库不同而存在差异，输入出错率高、工作效率低。检索内容涉及多个异构数据库时，无法实现跨数据库检索。只能通过终端检索信息，方式单一，在实际操作过程中存在误删和修改数据的安全隐患，急需一种异构信息检索的方法。

异构信息的存取有多种方法，目前有基于 Luene 开发 Solr、ELK、Kata 等开源检索工具，通过将信息分析和整理，建立统一的索引数据库用于数据的检索。但这些工具需要经过长期的专业学习才可熟练应用，且使用复杂，开发周期长[5,6]，并不适合 LAMOST 当前迫切的数据检索需求。结合实际，本文提出了一种异构环境信息快速检索方法。

本文首先简要介绍 LAMOST 环境信息的组成，然后在第 1 节介绍异构环境信息检索系统的组成和总体设计，系统各模块的实现以及整体系统的部署和测试在第 2、3 节进行详细的介绍，最后总结系统在 LAMOST 中的应用，并讨论 LAMOST 更多核心数据统一检索的可行性。

1 系统功能及设计

异构环境信息检索系统通过对异构数据库中环境信息的整理、归纳，屏蔽内部信息的复杂性，统一并简化检索语句，减少数据读取的出错概率。本系统的研发旨在保证原有数据安全的基础上，开发多种检索途径，快捷、精准获取数据，提高工作效率。

LAMOST 异构环境信息检索系统根据不同检索需求分为三部分：信息资源层、服务层以及应用层，如图 2。

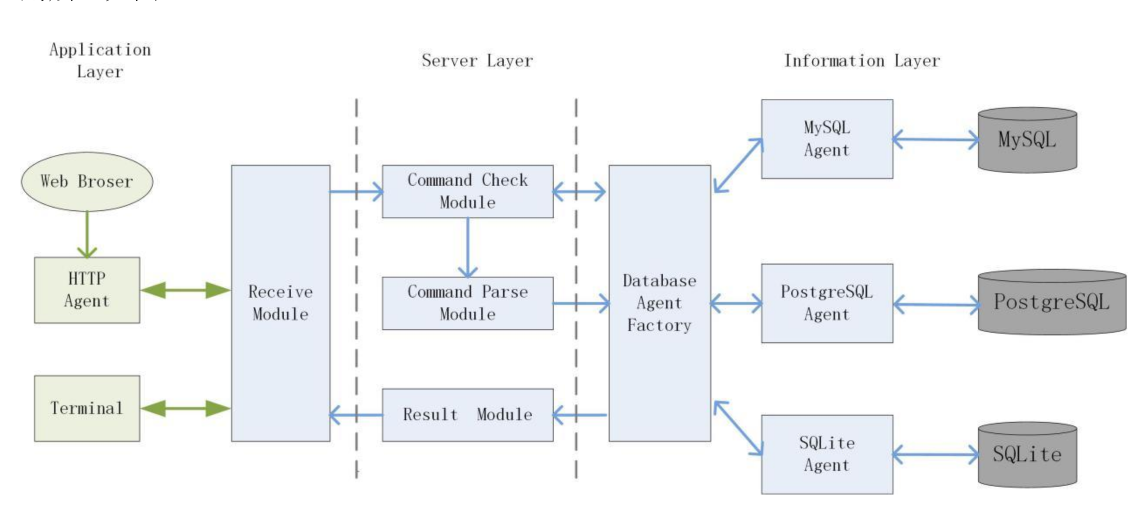


图 2 系统架构图

Figure 2 The system architecture

信息资源层包含数据库代理对象工厂模块和数据库异步代理模块。数据库代理对象工厂模块负责维护现有的数据库以及可检索信息，数据库异步代理模块负责 MySQL、PostgreSQL 等多种数据库的代理。

应用层为环境信息检索请求者，分为终端客户端和 web 客户端两种。终端客户端通过命令行检索，web 客户端通过 web 浏览器检索和展示结果。

服务层为环境信息检索服务器，用于接收客户端的检索请求，并根据请求内容转换为指定数据库类型的检索语法，最后将整合后的结果返回请求者。

2 系统实现

在研发过程中，根据功能分工，系统划分为数据库代理工厂模块、接收模块、命令检查模块、命令解析模块、结果汇总模块以及应用模块。各模块间相互解耦，简化了逻辑控制。

2.1 系统研发及运行环境

LAMOST 异构环境检索系统在 Python 3.7 开发环境下进行研发，可在多种 Linux 系列操作系统上运行。经过多方调研，根据实际工程需求，我们主要采用以下 Python 第三方资源库：

Asyncio：python3.7 的异步协程标准库，为大量第三方资源库提供异步协程支持。协程是由程序开发人员自主调度和控制，在用户态执行的一段代码，具有极高的执行效率，非常适合高并发型的网络通信[7-8]。

Aiomysql，aiopg，aiosqlit：Python3.7 中三个支持数据库异步操作的资源库，分别用于支持 MySQL、PostgreSQL、SQLite 数据库的异步操作。

Pyzmq:ZeroMQ 在 Python 中的开发库。作为一种高性能开源的消息中间件，ZeroMQ 具有丰富的通信资源库，且多种提供多种通信模式，可满足多种的通信业务需求[9]。

2.2 接收模块

接收模块是一个异步协程服务器，负责接收客户端的检索请求，并将检索结果返回检索请求者。

根据请求来源，检索分为望远镜内网、国家天文台内网及广域互联网三种方式，其中望远镜内网检索是最主要的工作方式。为确保系统提供稳定的环境信息，系统对外网检索请求数量进行了限制，一旦检索请求到达限制数量，则拒绝新的请求连接。

2.3 数据模块

数据模块分为数据库代理对象工厂模块和异步数据代理模块两部分。

LAMOST 观测和维护对环境数据检索需求基本固定，综合常见检索需求，系统设置了数据库代理对象工厂模块。如图 3 所示，该模块为自定义虚拟数据库，对当前所有数据库进行编号，以自定义检索数据名为关键字，映射检索数据相对应的数据库及字段名称。

数据库异步代理模块是数据库在系统中的代理，负责具体的数据检索和结果接收。根据数据库类型，数据库异步代理模块分为 MySQL 异步代理、PostgreSQL 异步代理以及 SQLite 异步代理。为提高数据库利用率，每个代理模块各维护一个有多个数据库连接的连接池，确保多个客户端的实时检索。

```
[DB]

sql1_pool_minsize = 1
sql1_pool_maxsize = 5
sql1_host = [redacted]
sql1_user =[redacted]
sql1_password =[redacted]
sql1_db =[redacted]

sql2_pool_minsize = 1
sql2_pool_maxsize = 5
sql2_host = [redacted]
sql2_user =[redacted]
sql2_password =[redacted]
sql2_db = ass

sql3_pool_minsize = 1
sql3_pool_maxsize = 5
sql3_host = [redacted]
sql3_user =[redacted]
sql3_password =[redacted]
sql3_db = [redacted]

sql4_pool_minsize = 1
sql4_pool_maxsize = 5
sql4_host = [redacted]
sql4_user =[redacted]
sql4_password =[redacted]
```

```
[GUIDE2_SEEING_INFO]
ccd2_seeing = fwhm, datatime, ccd2imagesfwhm, assimagesfwhmnew, 2

[GUIDE3_SEEING_INFO]
ccd3_seeing = fwhm, datatime, ccd3imagesfwhm, assimagesfwhmnew, 2

[FP_POSITION_INFO]
fp_pos = fppos, datatime, fp_run, ass, 2

[AO_INFO]
ao = value, timestamp, wavefrontdata, active_optic, 2

[3D_WIND_INFO]
3d_windspeed_avg1 = avg_ws, date, avg_data, wind_sonic, 1
3d_windspeed_avg2 = avg_ws, date, avg_data, wind_sonic, 1

[WEATHER_INFO]
temperature = temp, time, weather_data_new, ambient, 1
pressure = pres, time, weather_data_new, ambient, 1
humidity = rhum, time, weather_data_new, ambient, 1
dewpoint = dewpoint, time, weather_data_new, ambient, 1
windspeed_avg = winds_avg, time, weather_data_new, ambient, 1,3
windspeed_min = winds_min, time, weather_data_new, ambient, 1,3
windspeed_max = winds_max, time, weather_data_new, ambient, 1,3
winddirection = windd_avg, time, weather_data_new, ambient, 1,3

[GUIDE_MAX_MAY_FP_INFO]
guide_max = maX, obsdate, obstime, asscal_ma_fp, ass, 2
guide_may = maY, obsdate, obstime, asscal_ma_fp, ass, 2
```

图 3 数据库和命令映射

Figure 3 The map of database and command

2.4 命令检查模块

命令检查模块负责命令格式、内容和权限的审核。由于数据库检索需求多样化、检索者身份权限也不相同，因此数据库检索的便携性和安全性尤为重要。表 1 为自定义简化检索命令格式，只有符合格式的命令才可进行下一步的操作。为提高数据安全性，本系统对用户提交的检索请求进行安全性检查，仅允许对数据库采取检索操作。

表 1 自定义检索命令格式

Table 1 Custom search command format

Command format	Note
select <name>	Data within 24 hours
select <name> <new>	The newest data
select <name> <start time>	Data from start time to current time
select <name> <start time> <end time>	Data from start time to end time

2.5 命令解析模块

命令解析模块负责对检索命令的解析、转换及异步数据库代理的选择。数据库代理对象工厂模块中没有具体的望远镜运行环境信息。根据检索内容，解析模块在数据库代理对象工厂模块中获取数据所在的数据库信息及对应的字段名，重新构造检索语句后，根据数据库信

息选择异步数据库代理，将检索语句提交给代理进行具体检索操作。

2.6 结果汇总模块

结果汇总模块负责对检索结果的整理和转换。数据库异步代理模块将检索结果返回后，结果汇总模块会将其转化为统一的自定义结构，便于应用层数据的接收和整理。

2.7 应用模块

应用模块主要用于检索命令的提交和检索结果的展示，根据作用可分为命令客户端和 web 客户端两种方式。

LAMOST 内部有数百台计算机，操作系统种类繁多，为每台计算机安装一个客户端，工作量非常大，不合实际。由于工作人员在望远镜维护中只需检索环境数据，无需进一步的分析和记录，我们利用 Linux 和 MacOS 操作系统自带的 NetCat 工具包实现运行环境数据的检索。Windows 系统可以通过网络下载 NetCat 工具包，经过简单的系统设置即可使用。图 4 为 NetCat 用命令行方式检索运行环境信息。

HTTP 代理模块为外部检索访问提供 web 服务，我们使用 apache httpd 作为 web 服务器。如图 5 所示，http 代理接受浏览器发送的 http 请求后，将其转化成自定义检索命令并发送至服务层，并将服务层的检索结果转化成 http 后返回给浏览器，由浏览器展现。

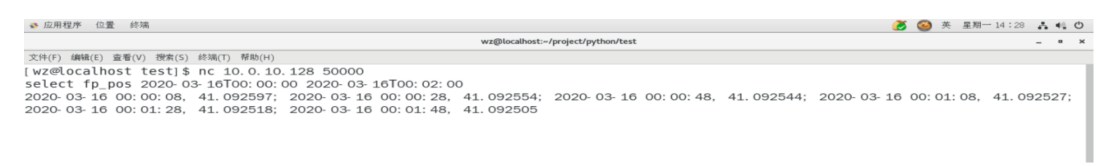


图 4 命令客户端检索

Figure 4 Search data through command client

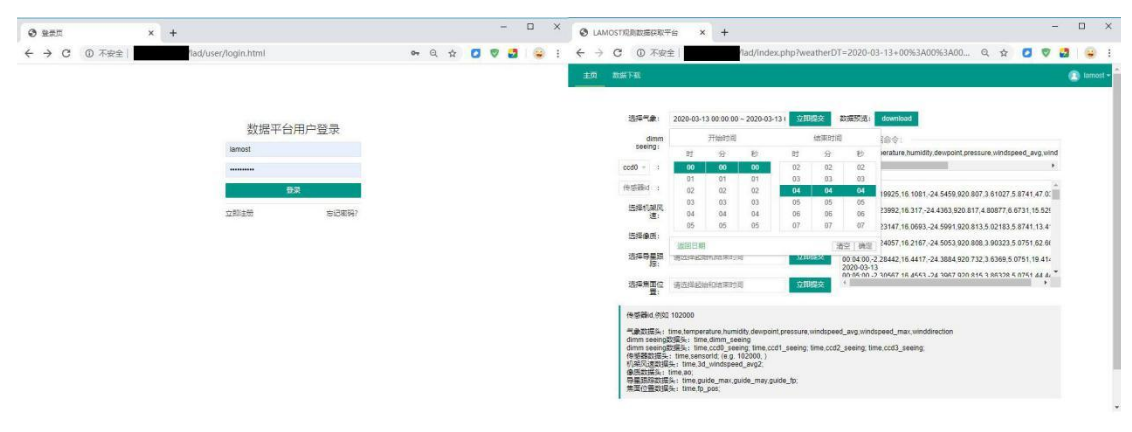


图 5 Web 检索界面

Figure 5 Search data though Web client

3 系统测试

为验证该系统是否满足 LAMOST 运行环境信息的检索需求，我们将其部署于 LAMOST 内一台多网段服务器上，为不同网域提供检索服务。图 7 为系统部署网络拓扑图，该系统可同时向 LAMOST 内网、国家天文台内网以及外部互联网提供检索服务。

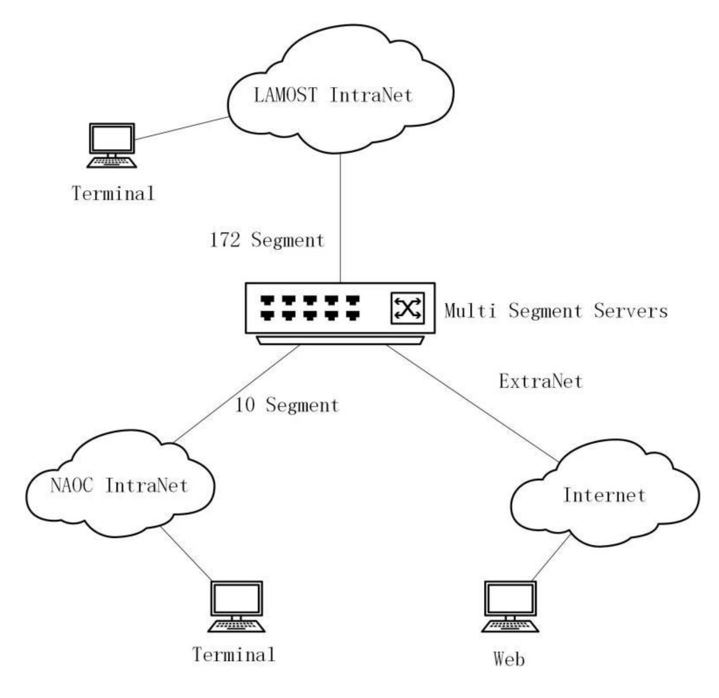


图 6 系统部署网络拓扑图

Figure 6 The network topology of system

多网段服务器配置为 Intel Xeon 处理器（4 个 CPU，每个 14 核），128G 内存，同数据库一起部署在 LAMOST 望远镜内部。LAMOST 内网测试电脑配置为 Intel Core i3-2100(双核)，4G 内存。国家天文台内网和外网测试电脑配置为 Intel Core i7-4790（双核）,16G 内存。由于该服务器承担着其他网络传输及数据处理任务，为了不影响正常使用，每网络段由 100 个客户端同时检索，每个客户端执行 3000 次检索请求。

如图 7 所示，检索 30 万次，172 网段总用时 9.936s，平均耗时 0.033ms；10 网段总用时 14.448s，平均耗时 0.048ms；外网总用时 21.155s，平均耗时 0.071ms。服务器在此期间 CPU 占用率为 43.5%,内存消耗可忽略不计。虽然外网平均检索用时最多，但是 0.071ms 的检索速度完全满足工作需求，服务器资源消耗低，不影响其他应用的正常运行。

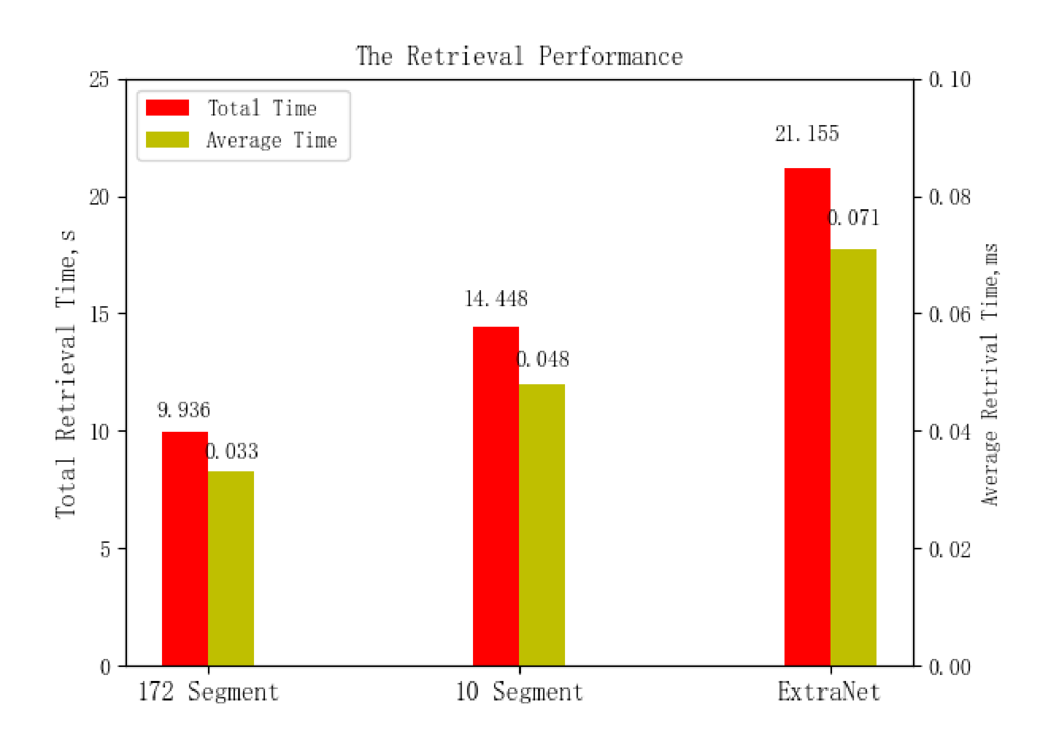


图 8 三网段检索性能测试

Figure 7 Search performance of three network segments

4 总结和展望

本系统采用了最新 python 3.7 的异步协程特性，降低了信息的获取难度，提高了系统吞吐能力；通过数据库代理对象工厂模式，整合多种数据库，增强了系统的可扩展性；使用自定义检索命令的方式简化了用户检索接口，提高了友好性，方便工作人员及时掌握望远镜运行环境信息；使用命令筛查制度，保护大型望远镜内部运行时细节不暴露，提高服务精度并维护数据的安全性。

本系统为将来对 LAMOST 所有数据库统一管理提供了基础框架，并设置了多种数据库接口，方便信息资源的后期扩展。

参考文献:

[1] 赵永恒.天文望远镜的自动观测技术[J].科研信息化技术与应用,2012,3(4):11-16.
Zhao Yongheng.Technology of automatic observation of astronomical telescope [J].E-Science Technology & Application,2012,3(4):11-16.

[2] 罗阿理,田园,宋静,等.LAMOST 观测控制系统设计与实现[J].科研信息细化技术与应用,2012,3(4):76-85.
Luo Ali，Tian Yuan，Song Jing，et al.Design and implementation of LAMOST observatory control system [J]. E-Science Technology & Application，

2012,3(4) :76－85.

[3] 李为民,邢晓正,胡红专,等. LAMOST 天文望远镜球焦面多光纤并行控制系统研究 [J]. 机械工程学报,2002,38(7):116-120.

Li Weimin,Xing Xiaozheng,Hu Hongzhuan,et al.Reasearch on the parallel control system of the multioptical fiber on the spherical focal plate for LAMOST [J].Chinese journal of mechanical engineering,2002,38(7):116-120.

[4] 李璋琪. 基于异构数据库的历史数据中心建设[J]. 电子技术于软件工程,2019,(18):154-158.

[5] 印奇,李青,黄鹏.基于 Solr 的飞机故障异构信息检索系统设计[J].航空科学技术,2017,28(04):30-36.

[6] 潘春华.基于 ELK 的高校数据中心运维多层面监控平台研究与实践[J].中国教育信息化,2020,(07),93-96.

[7] 刘健,黄才胜.基于协程的高并发架构研究[J].数字技术于应用,2018,36(04):85-86.

Liu Jian,Huang Caisheng.Research on high concurrency architecture based on CO process [J].Digital technology & application,2018,36(04):85-86.

[8] 田园,王锋,李建等.基于 Python 协程技术的 LAMOST 控制节点分布状态采集与监视系统[J].天文研究与技术,2018,15(4):456-464.

Tian Yuan,Wang Feng,Li Jian,et al.LAMOST resource monitoring system based on python coroutines technique[J].Astronomical research and technology,2018,15(4):456-464.

[9] 邓辉,钟文杰,付映雪等.基于 ZeroMQ 的新一代望远镜自动控制系统的通信框架设计[J],天文研究与技术,2018,15(03),308-314.

Deng Hui,Zhong Wenjie,Fu Yingxue,etl.The Design of Communication Framework for a New Generation of Telescope Automous Control System Based on ZeroMQ[J],Astronomical research and technology,2018,15(03),308-314.

Design and Implementation of Heterogeneous Environment Data Retrieval system for LAMOST

Wang Zheng[1,3],Wang Feng[2],Tian Yuan[1],Li Jian[1],Zhao Yongheng[1,3]

(1. Key Laboratory of Optical Astronomy,National Astronomical Observatories,Chinese Academy of Sciences,Beijing 100101,China,Email:wzheng@bao.ac.cn;

2.Astrophysics Center,Guangzhou University,Guangzhou 510006,China;

3.University of Chinese Academy of Sciences,Beijing 100049,China)

Abstract: The environment data is an important part of Large Sky Area Multi-Object Fiber Spectroscopy Telescope (LAMOST) for daily maintenance of the telescope,observation and the offline data processing. The environmental information is obtained from different subsystems designed by different institutes. The environment data storage formats are different and recorded in different databases, which makes them difficult to retrieve through a consistent interface. In this work, we designed a asynchronous coroutines system for the environment data retrieval from the LAMOST heterogeneous databases. This system creates a database agent factory module to collect the base information of the environment databases and the retrieve keywords of the environment data. This advantage makes it easier to composite the content of the retrieval data for different users and more convenient to upgrade the databases in the future.In last summer, by using the system in the operation and maintenance of LAMOST,users can retrieve the environment data via a client with custom commands or a web browser interface,that makes the operation of data retrieve of environmental information is easier;the content of the retrieval data is richer;the result of the data is more accurate,and finally improved the maintenance and observation efficiency of the telescope.

Key words: Data Retrieve;Asynchronous coroutines;LAMOST;